Tutorial is limited to pyspark transition to pandas users. Databricks would be seperate

In [1]:

```
import pandas as pd
from pyspark.sql import SparkSession
from pyspark import SparkContext
spark = SparkSession.builder.appName('SparkApp').getOrCreate()
```

Pyspark includes

- Native query

Pyspark doesn't include

- Indices

Hadoop Consistes of 3 layers

- Hdfs : file storage
- Spark : for processing and manipulation of data (replacement of mapreduce)
- Yarn : resource management

In [55]:

```
from pyspark.sql import functions as f
import numpy as np
```

In [8]:

```
# read file in pyspark

df_spark = spark.read.options(header='True', inferSchema='True', delimiter=',') \
  .csv("test1.csv")

#header = true - assumes header - first row
#inferSchema - detects data type
#delimiter - specifying delimiter

df_spark.show(1)
```

```
+-----+---+----------+------+
| Name|age|Experience|Salary|
+-----+---+----------+------+
|Krish| 31|        10| 30000|
+-----+---+----------+------+
only showing top 1 row
```

In [9]:

```
# read file in pandas

df_pandas = pd.read_csv('test1.csv', sep = ',')
df_pandas.head(1)
```

Out[9]:

	Name	age	Experience	Salary
0	Krish	31	10	30000

In [4]:

```
# writing in pyspark

#df_spark.write.options(header='True', delimiter=',').csv("output")
#df_spark.write.format("csv").save("output")
#df_spark.write.csv("test.csv")
#.mode('overwrite')
#df_spark.coalesce(1).write.csv('result.csv')
#df_spark.coalesce(1).write.csv("header.csv", header="true")


#df.write() API will create multiple part files inside given path
#to force spark write only a single part file use df.coalesce(1).write.csv(...)
#instead of df.repartition(1).write.csv(...) as coalesce is a narrow transformation whe
reas repartition is a wide transformation
```

In [6]:

```
# writing in pandas

df.to_csv('file.csv', index=None)
```

In [12]:

```
## Convert to pandas from spark
pd_df = df_spark.toPandas()
## Convert into Spark DataFrame
spark_df = spark.createDataFrame(pd_df)
```

In [ ]:

```
#spark save as delta table
spark_df.write.mode("overwrite").saveAsTable("temp.eehara_trial_table_9_5_19")

#you can create a new pandas dataframe witht the following command:
pd_df = spark.sql('select * from temp.eehara_trial_table_9_5_19').toPandas()

#spark query the delta table----------------------------------------

%sql

select * from temp.eehara_trial_table_9_5_19 ;
```

In [13]:

```
# view dataframe

#pandas

df_pandas.head(10)  # first 10
df_pandas.tail(5) #last 10

#spark

df_spark.show(1) #shows 1 rows
df_spark.limit(1) #limit 1 first rows
#df_spark.display() # to download - specific to databrics
```

```
+-----+---+----------+------+
| Name|age|Experience|Salary|
+-----+---+----------+------+
|Krish| 31|        10| 30000|
+-----+---+----------+------+
only showing top 1 row
```

Out[13]:

```
DataFrame[Name: string, age: int, Experience: int, Salary: int]
```

In [14]:

```
# pandas columns and data types

df_pandas.columns
df_pandas.dtypes

# pyspark columns and data types

df_spark.columns
df_spark.dtypes
```

Out[14]:

```
[('Name', 'string'), ('age', 'int'), ('Experience', 'int'), ('Salary', 'in
t')]
```

Renaming columns

In [17]:

```
#Pandas

df_pandas.rename(columns = {"Experience":"Exp"})

#pyspark

df_spark.withColumnRenamed("Experience","Exp")
```

Out[17]:

```
DataFrame[Name: string, age: int, Exp: int, Salary: int]
```

Dropping columns

In [19]:

```
#pandas

df_pandas.drop('Experience', axis = 1)

#spark

df_spark.drop('Experience')
```

Out[19]:

```
DataFrame[Name: string, age: int, Salary: int]
```

Filtering

In [22]:

```
# pandas

df_pandas[df_pandas['age'] > 10]
df_pandas[(df_pandas['age'] > 20) & (df_pandas['Experience'] > 5)]

#spark

df_spark[df_spark['age'] > 10]
df_spark[(df_spark['age'] > 20) & (df_spark['Experience'] > 5)]
```

Out[22]:

```
DataFrame[Name: string, age: int, Experience: int, Salary: int]
```

In [29]:

```
#in pyspark you can also do

### Salary of the people less than or equal to 20000
#df_spark.filter("Experience<=20").show()

#and

df_spark.filter(~(df_spark['Salary']<=27000) |
                  (df_spark['age']>= 31)).show()
```

```
+-----+---+----------+------+
| Name|age|Experience|Salary|
+-----+---+----------+------+
|Krish| 31|        10| 30000|
+-----+---+----------+------+
```

adding column

In [30]:

```
# pandas

df_pandas['retirement age'] = df_pandas['age'] + 3

#pyspark

df_spark.withColumn('retirement age', df_spark.age + 3)
```

Out[30]:

```
DataFrame[Name: string, age: int, Experience: int, Salary: int, retirement
age: int]
```

In [31]:

```
# filling nulls # can be used for whole df

# pandas

df_pandas['age'].fillna(0)

#spark

df_spark['age'].fillna(0)  #oops
```

```
---------------------------------------------------------------------------
-
TypeError                                 Traceback (most recent call las
t)
<ipython-input-31-c34e12d40add> in <module>
      7 #spark
      8
----> 9 df_spark['age'].fillna(0)

TypeError: 'Column' object is not callable
```

In [36]:

```
# using select function from pyspark

df_spark.select(f.col('age')).fillna(0)
```

Out[36]:

```
DataFrame[age: int]
```

group bys

In [40]:

```
df_pandas.groupby(['Name']).agg({'age':'max'})

#or

#df_pandas.groupby(['Name'], as_index = False)['age'].max()

#which is same as

#df_pandas.groupby(['Name'])['age'].max().reset_index()
```

Out[40]:

Name	age
Harsha	21
Krish	31
Paul	24
Shubham	23
Sudhanshu	30
Sunny	29

In [42]:

```
#spark

df_spark.groupby(['Name']).agg({'age':'max'}).show()
```

```
+---------+--------+
|     Name|max(age)|
+---------+--------+
|Sudhanshu|      30|
|    Sunny|      29|
|    Krish|      31|
|   Harsha|      21|
|     Paul|      24|
|  Shubham|      23|
+---------+--------+
```

renaming columns after groupby

In [50]:

```
#pandas

df_pandas.groupby(['Name'], as_index=False).agg({'age':'max'}).rename(columns = {"Exper
ience":"Exp"})

#to keep all columns you must specify aggregation for those columns too or group by the
m
```

Out[50]:

	Name	age
0	Harsha	21
1	Krish	31
2	Paul	24
3	Shubham	23
4	Sudhanshu	30
5	Sunny	29

In [53]:

```
#pyspark

df_spark.groupby(['Name']).agg({'age':'max'}).withColumnRenamed("max(age)","age").show
()
```

```
+---------+---+
|     Name|age|
+---------+---+
|Sudhanshu| 30|
|    Sunny| 29|
|    Krish| 31|
|   Harsha| 21|
|     Paul| 24|
|  Shubham| 23|
+---------+---+
```

conditionals statements

In [64]:

```
#pandas

df_pandas['eligible'] = np.where(df_pandas['age'] > 25, "Yes", "No")
df_pandas.head(2)
```

Out[64]:

	Name	age	Experience	Salary	retirement age	eligible
0	Krish	31	10	30000	34	Yes
1	Sudhanshu	30	8	25000	33	Yes

In [67]:

```
#pyspark

df_spark.withColumn('eligible', f.when(df_spark['age'] > 30, "yes").when(df_spark['Sala
ry'] > 0, "Yip").otherwise("No")).show(2)
```

```
+---------+---+----------+------+--------+
|     Name|age|Experience|Salary|eligible|
+---------+---+----------+------+--------+
|    Krish| 31|        10| 30000|     yes|
|Sudhanshu| 30|         8| 25000|     Yip|
+---------+---+----------+------+--------+
only showing top 2 rows
```

In [68]:

```
#there are bunch of ways in pandas for multiple conditions, below in one another exampl
e

##df = pd.DataFrame({'Type':list('ABBC'), 'Set':list('ZZXY')})
##conditions = [
##    (df['Set'] == 'Z') & (df['Type'] == 'A'),
##    (df['Set'] == 'Z') & (df['Type'] == 'B'),
##    (df['Type'] == 'B')]
##choices = ['yellow', 'blue', 'purple']
##df['color'] = np.select(conditions, choices, default='black')
##print(df)
```

lambda functions

In [69]:

```
#pandas

df_pandas['extra bonus'] = df_pandas['Salary'].apply(lambda x : x+ 3000)
df_pandas.head(2)
```

Out[69]:

	Name	age	Experience	Salary	retirement age	eligible	extra bonus
0	Krish	31	10	30000	34	Yes	33000
1	Sudhanshu	30	8	25000	33	Yes	28000

In [70]:

```
#pyspark

fnctnwa = f.udf(lambda x : x + 3000)
df_spark.withColumn('extra bonus',  fnctnwa(df_spark.Salary)).show(2)
```

```
+---------+---+----------+------+--------+-----------+
|     Name|age|Experience|Salary|eligible|extra bonus|
+---------+---+----------+------+--------+-----------+
|    Krish| 31|        10| 30000|     yes|      33000|
|Sudhanshu| 30|         8| 25000|     yes|      28000|
+---------+---+----------+------+--------+-----------+
only showing top 2 rows
```

JOINS

pyspark

In [71]:

```
df1=spark.read.csv('test1.csv',header=True,inferSchema=True).limit(2)
df1.show()
df2=spark.read.csv('test1.csv',header=True,inferSchema=True).limit(2)
df2.show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
+---------+---+----------+------+

+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
+---------+---+----------+------+
```

In [73]:

```
##1. when column name are same and you dont want duplicate column names

df1.join(df2, on='Name').show()
```

```
+---------+---+----------+------+---+----------+------+
|     Name|age|Experience|Salary|age|Experience|Salary|
+---------+---+----------+------+---+----------+------+
|    Krish| 31|        10| 30000| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000| 30|         8| 25000|
+---------+---+----------+------+---+----------+------+
```

In [75]:

```
# multiple columns , same name

df1.join(df2, on=['Name','age'], how='left').show()
```

```
+---------+---+----------+------+----------+------+
|     Name|age|Experience|Salary|Experience|Salary|
+---------+---+----------+------+----------+------+
|    Krish| 31|        10| 30000|        10| 30000|
|Sudhanshu| 30|         8| 25000|         8| 25000|
+---------+---+----------+------+----------+------+
```

In [76]:

```
#left on, right on for different column names wont work.. so need to rename column with
above technique

df2 = df2.withColumnRenamed('Name','Name2').withColumnRenamed('Experience','Experience
2')
df1.join(df2, ((df1.Name == df2.Name2) & (df1.Experience == df2.Experience2)), how = 'l
eft' ).show()
```

```
+---------+---+----------+------+---------+---+-----------+------+
|     Name|age|Experience|Salary|    Name2|age|Experience2|Salary|
+---------+---+----------+------+---------+---+-----------+------+
|    Krish| 31|        10| 30000|    Krish| 31|         10| 30000|
|Sudhanshu| 30|         8| 25000|Sudhanshu| 30|          8| 25000|
+---------+---+----------+------+---------+---+-----------+------+
```

pandas

In [77]:

```
df1=pd.read_csv('test1.csv')
df2=pd.read_csv('test1.csv')
df2.head(2)
```

Out[77]:

	Name	age	Experience	Salary
0	Krish	31	10	30000
1	Sudhanshu	30	8	25000

In [79]:

```
df1.merge(df2, left_on=['Name'], right_on=['Name'], how = 'left').head(2)
```

Out[79]:

	Name	age_x	Experience_x	Salary_x	age_y	Experience_y	Salary_y
0	Krish	31	10	30000	31	10	30000
1	Sudhanshu	30	8	25000	30	8	25000

In [80]:

```
#or

df1.merge(df2, on=['Name'], how = 'left').head(2)
```

Out[80]:

	Name	age_x	Experience_x	Salary_x	age_y	Experience_y	Salary_y
0	Krish	31	10	30000	31	10	30000
1	Sudhanshu	30	8	25000	30	8	25000

pivot table

In [81]:

```
#pandas

pd.pivot_table(df1, values='Salary',index=
               'Name', columns='age', aggfunc= np.sum)
```

Out[81]:

age	21	23	24	29	30	31
Name						
Harsha	15000.0	NaN	NaN	NaN	NaN	NaN
Krish	NaN	NaN	NaN	NaN	NaN	30000.0
Paul	NaN	NaN	20000.0	NaN	NaN	NaN
Shubham	NaN	18000.0	NaN	NaN	NaN	NaN
Sudhanshu	NaN	NaN	NaN	NaN	25000.0	NaN
Sunny	NaN	NaN	NaN	20000.0	NaN	NaN

In [83]:

```
#pyspark

df_spark.groupby('Name').pivot('age').sum('Salary').show()
```

```
+---------+-----+-----+-----+-----+-----+-----+
|     Name|   21|   23|   24|   29|   30|   31|
+---------+-----+-----+-----+-----+-----+-----+
|Sudhanshu| null| null| null| null|25000| null|
|    Sunny| null| null| null|20000| null| null|
|    Krish| null| null| null| null| null|30000|
|   Harsha|15000| null| null| null| null| null|
|     Paul| null| null|20000| null| null| null|
|  Shubham| null|18000| null| null| null| null|
+---------+-----+-----+-----+-----+-----+-----+
```

ISIN

In [87]:

```
#pandas

df_pandas[df_pandas['Name'].isin(['Sunny'])]
```

Out[87]:

	Name	age	Experience	Salary	retirement age	eligible	extra bonus
2	Sunny	29	4	20000	32	Yes	23000

In [93]:

```
#spark #opposite # same

df_spark[~df_spark['Name'].isin(['Sunny'])].show(2)
```

```
+---------+---+----------+------+--------+
|     Name|age|Experience|Salary|eligible|
+---------+---+----------+------+--------+
|    Krish| 31|        10| 30000|     yes|
|Sudhanshu| 30|         8| 25000|     yes|
+---------+---+----------+------+--------+
only showing top 2 rows
```

ISIN for vlookup -- passing columns

In [99]:

```
#pandas

df1=pd.read_csv('test1.csv')
df2=pd.read_csv('test1.csv')
df1[df1['Name'].isin(df2['Name'])].head(2)
```

Out[99]:

	Name	age	Experience	Salary
0	Krish	31	10	30000
1	Sudhanshu	30	8	25000

In [113]:

```
df1=spark.read.csv('test1.csv',header=True,inferSchema=True).limit(2)
df2=spark.read.csv('test1.csv',header=True,inferSchema=True).limit(2).show(2)
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
+---------+---+----------+------+
```

In [102]:

```
df1[df1['Name'].isin(df2['Name'])]
```

```
---------------------------------------------------------------------------
-
TypeError                                 Traceback (most recent call las
t)
<ipython-input-102-6e2d10c0fe8d> in <module>
----> 1 df1[df1['Name'].isin(df2['Name'])]

TypeError: 'NoneType' object is not subscriptable
```

In [127]:

```
#### we get error here

##soo

my_list = list(
    df2.select('Name').distinct().toPandas()['Name']
)


df1[df1['Name'].isin(my_list)].show()
```

```
+---------+---+----------+------+
|     Name|age|Experience|Salary|
+---------+---+----------+------+
|    Krish| 31|        10| 30000|
|Sudhanshu| 30|         8| 25000|
+---------+---+----------+------+
```

In [ ]:

In [ ]: